Vector-valued Distribution Regression – Keep It Simple and Consistent

Zoltán Szabó

Joint work with Bharath K. Sriperumbudur (PSU), Barnabás Póczos (CMU), Arthur Gretton (UCL)

CSML Reading Group
Department of Statistics
University of Oxford
May 1, 2015
The task

- Samples: $\{(x_i, y_i)\}_{i=1}^l$. Goal: $f(x_i) \approx y_i$, find $f \in \mathcal{H}$.

- Distribution regression:
  - $x_i$-s are distributions,
  - available only through samples: $\{x_{i,n}\}_{n=1}^{N_i}$.

  $\Rightarrow$ Training examples: labelled bags.
Example: aerosol prediction from satellite images

- Bag := pixels of a multispectral satellite image over an area.
- Label of a bag := aerosol value.

Engineered methods [Wang et al., 2012]: $100 \times \text{RMSE} = 7.5 - 8.5$.
- Using distribution regression?
Wider context

Context:
- machine learning: multi-instance learning,
- statistics: point estimation tasks (without analytical formula).

Applications:
- computer vision: image = collection of patch vectors,
- network analysis: group of people = bag of friendship graphs,
- natural language processing: corpus = bag of documents,
- time-series modelling: user = set of trial time-series.
Several algorithmic approaches

1. Parametric fit: Gaussian, MOG, exp. family
   [Jebara et al., 2004, Wang et al., 2009, Nielsen and Nock, 2012].

2. Kernelized Gaussian measures:
   [Jebara et al., 2004, Zhou and Chellappa, 2006].

3. (Positive definite) kernels:
   [Cuturi et al., 2005, Martins et al., 2009, Hein and Bousquet, 2005].


5. Set metrics: Hausdorff metric [Edgar, 1995]; variants
Theoretical guarantee?

- MIL dates back to [Haussler, 1999, Gärtner et al., 2002].

  1. compact Euclidean domain.
  2. output $= \mathbb{R}$. 
Kernel, RKHS

- \( k : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R} \) kernel on \( \mathcal{D} \), if
  - \( \exists \varphi : \mathcal{D} \rightarrow H(\text{ilbert space}) \) feature map,
  - \( k(a, b) = \langle \varphi(a), \varphi(b) \rangle_H \) (\( \forall a, b \in \mathcal{D} \)).

- Kernel examples: \( \mathcal{D} = \mathbb{R}^d \) (\( p > 0, \theta > 0 \))
  - \( k(a, b) = (\langle a, b \rangle + \theta)^p \): polynomial,
  - \( k(a, b) = e^{-\|a - b\|_2^2/(2\theta^2)} \): Gaussian,
  - \( k(a, b) = e^{-\theta\|a - b\|_2} \): Laplacian.

- In the \( H = H(k) \) RKHS (\( \exists! \)): \( \varphi(u) = k(\cdot, u) \).
Kernel: example domains ($\mathcal{D}$)

- Euclidean space: $\mathcal{D} = \mathbb{R}^d$.
- Graphs, texts, time series, dynamical systems.
- Distributions.
Def.: $k : \mathbb{D} \times \mathbb{D} \rightarrow \mathbb{R}$ kernel is universal if

- it is continuous,
- $H(k)$ is dense in $(C(\mathbb{D}), \| \cdot \|_\infty)$.

Examples: on compact subsets of $\mathbb{R}^d$

$$k(a, b) = e^{-\frac{\|a-b\|_2^2}{2\sigma^2}}, \quad (\sigma > 0)$$

$$k(a, b) = e^{\beta \langle a, b \rangle}, (\beta > 0), \text{ or more generally}$$

$$k(a, b) = f(\langle a, b \rangle), \quad f(x) = \sum_{n=0}^{\infty} a_n x^n \quad (\forall a_n > 0)$$
Problem formulation ($Y = \mathbb{R}$)

- Given:
  - labelled bags $\hat{z} = \{ (\hat{x}_i, y_i) \}_{i=1}^l$,
  - $i^{th}$ bag: $\hat{x}_i = \{ x_{i,1}, \ldots, x_{i,N} \} \sim x_i \in \mathcal{M}_1^+ (\mathcal{D}), \ y_i \in \mathbb{R}$.
- Task: find a $\mathcal{M}_1^+ (\mathcal{D}) \rightarrow \mathbb{R}$ mapping based on $\hat{z}$.
Problem formulation ($Y = \mathbb{R}$)

- Given:
  - labelled bags $\hat{\mathcal{Z}} = \{(\hat{x}_i, y_i)\}_{i=1}^l$,
  - $i^{th}$ bag: $\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \sim x_i \in \mathcal{M}_1^+ (\mathcal{D}), y_i \in \mathbb{R}$.
- Task: find a $\mathcal{M}_1^+ (\mathcal{D}) \rightarrow \mathbb{R}$ mapping based on $\hat{\mathcal{Z}}$.
- Construction: distribution embedding ($\mu_x$) + ridge regression

$$
\mathcal{M}_1^+ (\mathcal{D}) \xrightarrow{\mu=\mu(k)} X \subseteq H = H(k) \xrightarrow{f \in \mathcal{H} = \mathcal{H}(K)} \mathbb{R}.
$$
Problem formulation \((Y = \mathbb{R})\)

- **Given:**
  - labelled bags \(\hat{z} = \{(\hat{x}_i, y_i)\}^l_{i=1}\),
  - \(i^{th}\) bag: \(\hat{x}_i = \{x_{i,1}, \ldots, x_{i,N}\} \sim_{i.d.} x_i \in \mathcal{M}_1^+(\mathcal{D}),\ y_i \in \mathbb{R}\).

- **Task:** find a \(\mathcal{M}_1^+(\mathcal{D}) \to \mathbb{R}\) mapping based on \(\hat{z}\).

- **Construction:** distribution embedding \((\mu_x) + \text{ridge regression}\)

\[
\mathcal{M}_1^+(\mathcal{D}) \xrightarrow{\mu=\mu(k)} X \subseteq H = H(k) \xrightarrow{f \in \mathcal{H}=\mathcal{H}(K)} \mathbb{R}.
\]

- **Our goal:** risk bound compared to the regression function

\[
f_\rho(\mu_x) = \int_{\mathbb{R}} y d\rho(y | \mu_x).
\]
Goal in details

Contribution: analysis of the excess risk

\[ \mathcal{E}(f_{\hat{\lambda}}^\lambda, f_\rho) = \mathcal{R}[f_{\hat{\lambda}}^\lambda] - \mathcal{R}[f_\rho] \leq g(l, N, \lambda) \to 0 \text{ and rates,} \]

\[ \mathcal{R}[f] = \mathbb{E}_{(x,y)} |f(\mu_x) - y|^2 \text{ (expected risk),} \]

\[ f_{\hat{\lambda}}^\lambda = \arg\min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} |f(\mu_{\hat{x}_i}) - y_i|^2 + \lambda \|f\|_{\mathcal{H}}^2, \quad (\lambda > 0). \]

We consider two settings:

1. well-specified case: \( f_\rho \in \mathcal{H}, \)
2. misspecified case: \( f_\rho \in L^2_{\rho x} \setminus \mathcal{H}. \)
Step-1: mean embedding

- $k : \mathcal{D} \times \mathcal{D} \rightarrow \mathbb{R}$ kernel; canonical feature map: $\varphi(u) = k(\cdot, u)$.
- Mean embedding of a distribution $x$, $\hat{x}_i \in \mathcal{M}_1^+(\mathcal{D})$:

$$
\mu_x = \int_{\mathcal{D}} k(\cdot, u) dx(u) \in H(k),
$$

$$
\mu_{\hat{x}_i} = \int_{\mathcal{D}} k(\cdot, u) d\hat{x}_i(u) = \frac{1}{N} \sum_{n=1}^{N} k(\cdot, x_{i,n}).
$$

- Linear $K \Rightarrow$ set kernel:

$$
K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j}) = \langle \mu_{\hat{x}_i}, \mu_{\hat{x}_j} \rangle_H = \frac{1}{N^2} \sum_{n,m=1}^{N} k(x_{i,n}, x_{j,m}).
$$
Step-2: ridge regression (analytical solution)

Given:
- training sample: \( \hat{z} \),
- test distribution: \( t \).

Prediction:

\[
(f_\hat{z}^\lambda \circ \mu)(t) = k(K + l\lambda I_l)^{-1}[y_1; \ldots; y_l],
\]

\[
K = [K(\mu_{\hat{x}_i}, \mu_{\hat{x}_j})] \in \mathbb{R}^{l \times l},
\]

\[
k = [K(\mu_{\hat{x}_1}, \mu_t), \ldots, K(\mu_{\hat{x}_l}, \mu_t)] \in \mathbb{R}^{1 \times l}.
\]
Blanket assumptions

- $\mathcal{D}$: separable, topological domain.
- $k$:
  - bounded: $\sup_{u \in \mathcal{D}} k(u, u) \leq B_k \in (0, \infty)$,
  - continuous.
- $K$: bounded; Hölder continuous: $\exists L > 0, h \in (0, 1]$ such that
  \[
  \|K(\cdot, \mu_a) - K(\cdot, \mu_b)\|_{\mathcal{H}} \leq L \|\mu_a - \mu_b\|_H^h.
  \]
- $y$: bounded.
- $X = \mu\left(\mathcal{M}_1^+(\mathcal{D})\right) \in \mathcal{B}(H)$. 
Performance guarantees (in human-readable format)

If in addition

1. well-specified case: $f_\rho$ is 'c-smooth' with 'b-decaying covariance operator' and $l \geq \lambda^{-\frac{1}{b}}$, then

$$
\mathcal{E}(f_{2}^{\lambda}, f_{\rho}) \leq \frac{\log^h(l)}{N^h \lambda^3} + \lambda^c + \frac{1}{l^2 \lambda} + \frac{1}{l \lambda^{\frac{1}{b}}}. \tag{4}
$$

2. misspecified case: $f_\rho$ is 's-smooth', $L^2_{\rho_x}$ is separable, and $\frac{1}{\lambda^2} \leq l$, then

$$
\mathcal{E}(f_{2}^{\lambda}, f_{\rho}) \leq \frac{\log^\frac{h}{2}(l)}{N^\frac{h}{2} \lambda^{\frac{3}{2}}} + \frac{1}{\sqrt{l} \lambda} + \frac{\sqrt{\lambda^{\min(1,s)}}}{\lambda \sqrt{l}} + \lambda^{\min(1,s)}. \tag{5}
$$
Misspecified case: assume

- $s \geq 1$, $h = 1$ ($K$: Lipschitz),
- $1 = 3$ in (5) $\Rightarrow \lambda$; $l = N^a$ ($a > 0$)
- $t = lN$: total number of samples processed.

Then

1. $s = 1$ (’most difficult’ task): $\mathcal{E}(f^{\lambda}_z, f_\rho) \approx t^{-0.25}$,
2. $s \rightarrow \infty$ (’simplest’ problem): $\mathcal{E}(f^{\lambda}_z, f_\rho) \approx t^{-0.5}$. 

Zoltán Szabó

Vector-valued Distribution Regression
k: bounded, continuous $\Rightarrow$

- $\mu : (M_1^+(D), B(\tau_w)) \to (H, B(H))$ measurable.
- $\mu$ measurable, $X \in B(H) \Rightarrow \rho$ on $X \times Y$: well-defined.

If (*) := $D$ is compact metric, $k$ is universal, then

- $\mu$ is continuous, and
- $X \in B(H)$. 

Zoltán Szabó

Vector-valued Distribution Regression
In case of (*):

<table>
<thead>
<tr>
<th></th>
<th>$K_G$</th>
<th>$K_e$</th>
<th>$K_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = 1$</td>
<td>$e^{-\frac{|\mu_a - \mu_b|^2_H}{2\theta^2}}$</td>
<td>$e^{-\frac{|\mu_a - \mu_b|_H}{2\theta^2}}$</td>
<td>$(1 + |\mu_a - \mu_b|^2_H / \theta^2)^{-1}$</td>
</tr>
<tr>
<td>$h = \frac{1}{2}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$K_t$</th>
<th>$K_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h = \frac{\theta}{2}$ ((\theta \leq 2))</td>
<td>$(1 + |\mu_a - \mu_b|_H^\theta)^{-1}$</td>
<td>$(|\mu_a - \mu_b|^2_H + \theta^2)^{-\frac{1}{2}}$</td>
</tr>
<tr>
<td>$h = 1$</td>
<td></td>
<td>$h = 1$</td>
</tr>
</tbody>
</table>

They are functions of $\|\mu_a - \mu_b\|_H \Rightarrow$ computation: similar to set kernel.
Notes on the assumptions: misspecified case

$L^2_{\rho_X}$: separable $\iff$ measure space with $d(A, B) = \rho_X(A \triangle B)$ is so [Thomson et al., 2008].
Vector-valued output: $Y = \text{separable Hilbert}$

- **Objective function:**

  $$f^\lambda_{\hat{z}} = \arg \min_{f \in \mathcal{H}} \frac{1}{l} \sum_{i=1}^{l} \| f(\mu_{\hat{x}_i}) - y_i \|_Y^2 + \lambda \| f \|_\mathcal{H}^2, \quad (\lambda > 0).$$

- $K(\mu_a, \mu_b) \in \mathcal{L}(Y)$: vector-valued RKHS.
Analytical solution: prediction on a new test distribution \((t)\)

\[
(f^\lambda_\mathcal{Z} \circ \mu)(t) = k(K + I\lambda I_I)^{-1}[y_1; \ldots; y_l],
\]

\[
K = [K(\mu_{\mathcal{X}_i}, \mu_{\mathcal{X}_j})] \in \mathcal{L}(Y)^{l \times l},
\]

\[
k = [K(\mu_{\mathcal{X}_1}, \mu_t), \ldots, K(\mu_{\mathcal{X}_l}, \mu_t)] \in \mathcal{L}(Y)^{1 \times l}.
\]

Specially: \(Y = \mathbb{R} \Rightarrow \mathcal{L}(Y) = \mathbb{R}; Y = \mathbb{R}^d \Rightarrow \mathcal{L}(Y) = \mathbb{R}^d.\)
Boundedness and Hölder continuity of \( K \):

1. **Boundedness:**
\[
\| K_{\mu_a} \|^2_{\text{HS}} = Tr \left( K_{\mu_a}^* K_{\mu_a} \right) \leq B_K \in (0, \infty), \quad (\forall \mu_a \in X).
\]

2. **Hölder continuity:** \( \exists L > 0, \ h \in (0, 1] \) such that
\[
\| K_{\mu_a} - K_{\mu_b} \|_{\mathcal{L}(\mathcal{Y}, \mathcal{H})} \leq L \| \mu_a - \mu_b \|_{\mathcal{H}}^h, \quad \forall (\mu_a, \mu_b) \in X \times X.
\]
Problem: learn the entropy of the 1\textsuperscript{st} coo. of (rotated) Gaussians.

Baseline: kernel smoothing based distribution regression (applying density estimation) $\Rightarrow$ DFDR.

Performance: RMSE boxplot over 25 random experiments.

Experience:
- more precise than the only theoretically justified method,
- by avoiding density estimation.
Supervised entropy learning: plots

RMSE: MERR=0.75, DFDR=2.02

- Entropy
- Rotation angle ($\beta$)
- RMSE

true
MERR
DFDR

Zoltán Szabó
Vector-valued Distribution Regression
Performance: $100 \times \text{RMSE}$.

Baseline [mixture model (EM)]: $7.5 - 8.5 \ (\pm 0.1 - 0.6)$.

Linear $K$:
- single: $7.91 \ (\pm 1.61)$.
- ensemble: $\mathbf{7.86} \ (\pm 1.71)$.

Nonlinear $K$:
- Single: $7.90 \ (\pm 1.63)$,
- Ensemble: $\mathbf{7.81} \ (\pm 1.64)$. 
Summary

Problem: distribution regression.

Literature: large number of heuristics.

Contribution:
- a simple ridge solution is consistent,
- specifically, the set kernel is so (15-year-old open question).

Simplified version \( Y = \mathbb{R}, f_\rho \in \mathcal{H} \):
- accepted at AISTATS-2015 (oral).

Zoltán Szabó
Vector-valued Distribution Regression
MERR code (ITE toolbox), complete analysis (submitted to JMLR):

https://bitbucket.org/szzoli/ite/

Closely related research directions (Bayesian world):
- \( \infty \)-dimensional exp. family fitting,
Thank you for the attention!

**Acknowledgments:** This work was supported by the Gatsby Charitable Foundation, and by NSF grants IIS1247658 and IIS1250350. The work was carried out while Bharath K. Sriperumbudur was a research fellow in the Statistical Laboratory, Department of Pure Mathematics and Mathematical Statistics at the University of Cambridge, UK.
Appendix: contents

- Topological definitions, separability.
- Exact prior definitions.
- Vector-valued RKHS.
- Hausdorff metric.
- Weak topology on $\mathcal{M}^+_1(\mathcal{D})$. 
Given: \( \mathcal{D} \neq \emptyset \) set.

\( \tau \subseteq 2^\mathcal{D} \) is called a topology on \( \mathcal{D} \) if:

1. \( \emptyset \in \tau, \mathcal{D} \in \tau \).
2. Finite intersection: \( O_1 \in \tau, O_2 \in \tau \Rightarrow O_1 \cap O_2 \in \tau \).
3. Arbitrary union: \( O_i \in \tau \ (i \in I) \Rightarrow \bigcup_{i \in I} O_i \in \tau \).

Then, \( (\mathcal{D}, \tau) \) is called a topological space; \( O \in \tau \): open sets.
Given: \((\mathcal{D}, \tau)\). \(A \subseteq \mathcal{D}\) is

- **closed** if \(\mathcal{D}\setminus A \in \tau\) (i.e., its complement is open),
- **compact** if for any family \((O_i)_{i \in I}\) of open sets with \(A \subseteq \bigcup_{i \in I} O_i\), \(\exists i_1, \ldots, i_n \in I\) with \(A \subseteq \bigcup_{j=1}^n O_{i_j}\).

**Closure** of \(A \subseteq \mathcal{D}\):  

\[
\bar{A} := \bigcap_{A \subseteq C \text{ closed in } \mathcal{D}} C.
\]

\((\mathcal{D}, \tau)\) is **separable** if \(\exists\) countable, dense subset of \(\mathcal{D}\).

Counterexample: \(L^\infty / L^\infty\).
Prior (well-specified case): $\rho \in \mathcal{P}(b, c)$

- Let the $T : \mathcal{H} \rightarrow \mathcal{H}$ covariance operator be
  \[
  T = \int_X K(\cdot, \mu_a)K^*(\cdot, \mu_a)d\rho_X(\mu_a)
  \]
  with eigenvalues $t_n$ ($n = 1, 2, \ldots$).

- Assumption: $\rho \in \mathcal{P}(b, c) = \text{set of distributions on } X \times Y$
  \begin{itemize}
    
    - $\alpha \leq n^b t_n \leq \beta$ ($\forall n \geq 1; \alpha > 0, \beta > 0$),
    
    - $\exists g \in \mathcal{H}$ such that $f_\rho = T^{c-1/2}g$ with $\|g\|_{\mathcal{H}}^2 \leq R$ ($R > 0$),
  \end{itemize}

where $b \in (1, \infty)$, $c \in [1, 2]$.

- Intuition: $1/b$ – effective input dimension, $c$ – smoothness of $f_\rho$.  

Zoltán Szabó  
Vector-valued Distribution Regression
Let $\tilde{T}$ be defined as:

$S^*_K : \mathcal{H} \rightarrow L^2_{\rho_X}$,

$S_K : L^2_{\rho_X} \rightarrow \mathcal{H}$,  \quad $(S_K g)(\mu_u) = \int_X K(\mu_u, \mu_t)g(\mu_t)d\rho_X(\mu_t)$,

$\tilde{T} = S^*_K S_K : L^2_{\rho_X} \rightarrow L^2_{\rho_X}$.

Our range space assumption on $\rho$: $f_\rho \in \text{Im}\left(\tilde{T}^s\right)$ for some $s \geq 0$. 
Definition:

- A $\mathcal{H} \subseteq Y^X$ Hilbert space of functions is RKHS if
  \[
  A_{\mu_x,y} : f \in \mathcal{H} \mapsto \langle y, f(\mu_x) \rangle_Y \in \mathbb{R}
  \]
  is continuous for $\forall \mu_x \in X, y \in Y$.
- $\Rightarrow$ The evaluation functional is continuous in every direction.
Riesz representation theorem $\Rightarrow \exists K(\mu_x | y) \in \mathcal{H}$:

$$\langle y, f(\mu_x) \rangle_Y = \langle K(\mu_x | y), f \rangle_{\mathcal{H}} \quad (\forall f \in \mathcal{H}). \quad (11)$$

$K(\mu_x | y)$: linear, bounded in $y \Rightarrow K(\mu_x | y) = K_{\mu_x}(y)$ with $K_{\mu_x} \in \mathcal{L}(Y, \mathcal{H})$. 

Zoltán Szabó

Vector-valued Distribution Regression
Vector-valued RKHS: $\mathcal{H} = \mathcal{H}(K)$ – continued

- Riesz representation theorem $\Rightarrow \exists K(\mu_x | y) \in \mathcal{H}$:
  \[
  \langle y, f(\mu_x) \rangle_Y = \langle K(\mu_x | y), f \rangle_{\mathcal{H}} \quad (\forall f \in \mathcal{H}).
  \]  
  (11)

- $K(\mu_x | y)$: linear, bounded in $y \Rightarrow K(\mu_x | y) = K_{\mu_x}(y)$ with $K_{\mu_x} \in \mathcal{L}(Y, \mathcal{H})$.

- $K$ construction:
  \[
  K(\mu_x, \mu_t)(y) = (K_{\mu_t}y)(\mu_x), \quad (\forall \mu_x, \mu_t \in X), \text{ i.e.,}
  \]
  \[
  K(\cdot, \mu_t)(y) = K_{\mu_t}y,
  \]  
  (12)

  \[
  \mathcal{H}(K) = \text{span}\{K_{\mu_t}y : \mu_t \in X, y \in Y\}.
  \]  
  (13)
Riesz representation theorem $\Rightarrow \exists K(\mu_x|y) \in \mathcal{H}$:

$$\langle y, f(\mu_x) \rangle_Y = \langle K(\mu_x|y), f \rangle_{\mathcal{H}} \quad (\forall f \in \mathcal{H}).$$ (11)

$K(\mu_x|y)$: linear, bounded in $y$ $\Rightarrow K(\mu_x|y) = K_{\mu_x}(y)$ with $K_{\mu_x} \in \mathcal{L}(Y, \mathcal{H})$.

$K$ construction:

$$K(\mu_x, \mu_t)(y) = (K_{\mu_t}y)(\mu_x), \quad (\forall \mu_x, \mu_t \in X), \quad \text{i.e.,}$$

$$K(\cdot, \mu_t)(y) = K_{\mu_t}y, \quad (12)$$

$$\mathcal{H}(K) = \overline{\text{span}}\{K_{\mu_t}y : \mu_t \in X, y \in Y\}. \quad (13)$$

Shortly: $K(\mu_x, \mu_t) \in \mathcal{L}(Y)$ generalizes $k(u, v) \in \mathbb{R}$.  

Zoltán Szabó 
Vector-valued Distribution Regression
1. \( K_i : X \times X \to \mathbb{R} \) kernels \((i = 1, \ldots, d)\). Diagonal kernel:

\[
K(\mu_a, \mu_b) = \text{diag}(K_1(\mu_a, \mu_b), \ldots, K_d(\mu_a, \mu_b)). \tag{14}
\]

2. Combination of \( D_j \) diagonal kernels \([D_j(\mu_a, \mu_b) \in \mathbb{R}^{r \times r}, A_j \in \mathbb{R}^{r \times d}]\):

\[
K(\mu_a, \mu_b) = \sum_{j=1}^{m} A_j^* D_j(\mu_a, \mu_b) A_j. \tag{15}
\]
Existing methods: set metric based algorithms

- Hausdorff metric [Edgar, 1995]:

\[ d_H(X, Y) = \max \left\{ \sup_{x \in X} \inf_{y \in Y} d(x, y), \sup_{y \in Y} \inf_{x \in X} d(x, y) \right\}. \]  \hspace{1cm} (16)

- Metric on compact sets of metric spaces \([(M, d); X, Y \subseteq M]\).

- 'Slight' problem: highly sensitive to outliers.
Def.: It is the weakest topology such that the mapping is continuous for all \( h \in C_b(D) \), where

\[
C_b(D) = \{(D, \tau) \to \mathbb{R} \text{ bounded, continuous functions}\}.
\]

In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 870–873.


Nonparametric divergence estimation with applications to machine learning on distributions.
In *Uncertainty in Artificial Intelligence (UAI)*, pages 599–608.

k-NN regression on functional data with incomplete observations.
In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

*Real Analysis*.
Prentice-Hall.

Closed-form Jensen-Rényi divergence for mixture of Gaussians and applications to group-wise shape registration.

Solving the multiple-instance problem: A lazy learning approach.

